MICRO-ECONOMIC DATA BANKS: PROBLEMS AND POTENTIAL

HAROLD W. WATTS*
University of Wisconsin

Recent and revolutionary advances in data processing and computing machinery, combined with expanding bodies of data and increasing numbers of analysts with basic quantitative skills, have led to the view that we are entering a new era of social analysis. There is also a new urgency to tackle the many tough social problems that can only be solved by analysis at the micro-unit level, which may well lead people to *need* such a new era in data collection whether or not it is actually round the corner.

There is not, in fact, very much evidence in the form of completed research that the vast potential created by these advances in computer technology is yet being exploited. From my own experience in this area, I have developed a view as to why this is so, and in this note indicate that there are very difficult and unsolved problems involved in harnessing these resources, and that these problems are peculiar to the collection, storage, and usage of micro-data.

Micro-data are collected from direct surveys of individual units rather than from the putting together of many subsets of secondary information into large-scale aggregates. And it is not at all unlikely that such direct data collection may be relatively more feasible in Latin America than in fully industrialized countries like the United States, because substitutes for such statistics are harder to come by and less reliable.

The remainder of this paper will be organized in the following way. After defining in more detail what I mean by "micro-data," I shall recount two specific episodes in which I have been involved because they illustrate the problems of such undertakings; then I shall draw a few conclusions and give my advice about what criteria should guide the setting up of generally usable systems to handle such data.

MICRO-DATA

Micro-economic data are here taken to be information pertaining to, and unique to, specific decision-making *units*. These may be individuals, families, firms, political units, and so on. The data may be cross-sectional—giving information on a unit's status at a point in time (or for one period of time), or they may provide information referring to (or collected at) several successive points in time. The full complexity of such data is reached when the information is collected for a series of points in time. A person is indivisible for these purposes; he is, however, born and he does die. And, the decision units of which he is a part can also change

* I would like to acknowledge here the editorial skills of Felicity Skidmore which made my disjointed thoughts into a paper. This paper was prepared while the author was Irving Fisher Research Professor at Yale University.

from one survey to another. This process—of birth, death, and mutation of multi-person units—is what makes it difficult to organize, store, and work with micro-data. Different analyses are likely to apply to different decision units or even different versions of what is nominally the same unit. Thus, choices of definition have to be made, and the questions of how units are to be matched, put together, followed from survey to survey, depend on these choices.

To be useful for analysis, collections of micro-data should provide input for research that is timely, and also responsive to important areas of uncertainty. There is now in operation computing and file-manipulating machinery that is enormously powerful and becoming steadily less costly. Operating systems and program libraries also reflect a high degree of development, and are still active areas for innovation. And there is a wide range of storage media—cards, tapes, disks, drums—and on the horizon are even more exotic and compact media. Finally, there is a growing inventory of data born of a recognition that many questions require detailed information on families or other decision-making units, including data on how variables for specific units have changed over time.

Why have these resources not been exploited more fully? Why have the theory and practice of social and economic systems been able to draw on them to such a limited extent? The answer may lie along the following lines. The organizational effort and the budget required to join all these components together into a working system are beyond the capacity of individual researchers. Such researchers are, therefore, led to ignore micro-data and devote their efforts to more traditional, heavily worked over, manageable sources of data.[1] At the other extreme, research groups with generous resources have been working toward generating the super-colossal type of micro-data bank that aims at building up to a level of generality which is almost if not completely impossible given the current state of the art.

The next two sections describe the problems encountered in two relatively modest micro-data gathering efforts, to lend some realism to the discussion of massive general-purpose data banks, and to give some idea of the magnitude of the problems which must be overcome before we can hope to operationalize such a concept.

The Survey of Economic Opportunity

When President Johnson's War on Poverty was declared, certain antipoverty government programs were initiated by the Office of Economic Opportunity (OEO). It did not take long for the research staff of the agency to realize that not very much was known about the characteristics of the poor in the United States, and even less about the impact that OEO's action programs might be having on those poor people. It was, therefore, decided to get new data on these questions by administering a survey to a large number of families (30,000). Low-income census tracts were sampled more than proportionally because of the central purpose of the survey. All the dwelling units were interviewed in early 1966 and a subset of them were interviewed again in early 1967, along with a new (independently drawn) subsample to make up the same total.

[1] F. Thomas Juster, "Microdata, Economic Research, and the Production of Economic Knowledge," *American Economic Review*, May 1970.

Since OEO did not have the machinery to undertake the survey themselves, they contracted with the Census Bureau to do it for them. However, in addition to providing up-to-date information on the ten-ongoing poverty programs, the SEO was also designed to provide a data base for more fundamental analytic studies of the social process that produces and perpetuates poverty. The instrument, therefore, included a broader set of household variables than had been traditional in Census surveys.

Although the interviews took place in 1966 and 1967, it is only within the past year that any volume of analytic work has been produced using these data, and the longitudinal subsample has not yet been exploited on a wide scale. Also, when the basic information on size and status of various parts of the poverty population were initially pulled together and made available they conflicted with other sources, producing inconsistencies that have yet to be satisfactorily and completely resolved. Why the three-year lag—which was totally unpredicted by the planners and was never recognized as inevitable even when the data were being processed?

The first data tapes were made available (from the initial, 1966 survey wave) by the Census Bureau in late Spring of 1967. This was much later than everyone had expected for the results of the first cross-section. Indeed there had been plans to use its results to guide the second wave administered in the first months of 1967.

The fielding and administering of the questionnaires caused no apparent problem; but reliable transcription of the data from the questionnaires into analyzable form proved intractable to a degree which was a complete surprise to the Census Bureau—hardly a novice at large-scale data collection.

The problem centered on the fact that the Census organization was geared to the ordinary operation of a multi-program, data-production system that was completely routinized. The adaptation of this system to a different task proved unexpectedly difficult even for experienced technicians. Most prominently, the variables which were unique to the Survey of Economic Opportunity required both new conceptual work and new computer-programming work before the data could be edited and checked, and before missing items could be accounted for and allocated.

In fact the Census Bureau divided the task—processing themselves the part which could use the existing routines for the Current Population Survey (C.P.S.), and subcontracting (to ARIES Corporation) the new or unique segments. Unfortunately the coded identifiers for individual families were not always unique so that it proved impossible to put the two segments back together for some of the households. This error was not discovered until the Fall of 1967 after a substantial amount of effort had been spent on further "data cleaning."

But there was a second major problem as well, connected with the problem of making data in unaggregated form available to researchers. Providing so-called "raw" data was not something the U.S. Census Bureau had done routinely or comfortably. They observe very high standards for all statistical products made available for general consumption. Their sense of responsibility may even be said to have developed to the point where, in their efforts to preclude all possibility of foolish or perverse interpretation of their statistics, they prevent interpretation of any kind. This instance proved no exception. They were extremely uneasy about

releasing micro-data even to OEO (which commissioned them) for fear of the multifarious uses to which they might conceivably be put.

Their discomfiture was enhanced by another dimension to the problem. Many of the analyses anticipated for the SEO data involved multi-variable regression and multi-variate analysis. Such processes, of course, produce results that are much more sensitive to data editing and allocation practices than are the tabulations traditionally produced by the Census Bureau. In other words, cross tabulations usually have open-ended categories, and these can contain an occasional wild error without appreciable effect upon any interpretation that might be placed on the central or modal segment. Not so with more sophisticated statistical tools.

It is certainly the case that there is no reason to "clean" data beyond the point of diminishing returns for tabular analysis if that is all you need. But, at the same time, any census bureau must hesitate to provide ammunition for challenges to its authority; and the possibility that the data might not be absolutely clean when released must have been quite threatening. Since that time, however, the Bureau has relaxed its stance on release of micro-data, and it is now possible to get non-disclosing copies of the Current Population Survey tapes.

In any case, when the data were turned over to OEO in May of 1967 they were still well short of the micro-analytic standards the OEO sponsors required. Consequently, further data cleaning was contracted to the Brookings Institution, who, along with Assist Corp., also spent at least twice as much time on the job as they had anticipated. And, they also, no doubt, relinquished the data before being fully satisfied. The Brookings-Assist data, now including both annual Surveys were, however, clean enough in OEO's opinion to be made available to researchers on September 3, 1969 along with voluminous (and clear and complete) documentation, describing in detail the data on the actual magnetic tapes.

Now the cause for delay shifted to the potential users. In order to facilitate access to the data, OEO contracted with the Data and Computation Center at the University of Wisconsin to be the repository, distributor and service agency for the SEO files.[2] Consultation and guidance were also to be provided by the Institute for Research on Poverty. At the same time several other, mostly university-based, researchers obtained copies of the tapes. But despite the excellence of documentation, all users experienced unexpected delays of from two to six months in getting the data "running," *i.e.* in gaining enough familiarity with the files so that at least half the attempts to use it were successful.

After the required familiarity had been established, however, the data file appeared to be unnecessarily costly to use. It was clear that a specified restructuring would literally decimate the costs of any analysis at Wisconsin. We could not afford to ignore such a large cost factor, so the restructuring took place, with further frustrating delays for researchers who had by now been anticipating being able to use the data for three years. They were finally able to begin their analyses in the summer of 1970.

[2] See the note by Max E. Ellis, "Social Science Computing at the University of Wisconsin: SIMS and SEOSYS," *Annals of Economic and Social Measurement*, Vol. 1, Number 2, April 1972.

Such work as has been done utilizes mainly the cross-sectional aspect of the SEO. So far very little work has been done with the continuous data records from both years. And there still remain further problems for users when the longitudinal aspects of the data begin to be exploited on a wider scale.

Two major problems exist. First, the longitudinal property of the data lies in the fact that the same "dwelling unit" was interviewed each time. Obviously this means that the same family may or may not have been there the second time. A certain number of records, therefore, are not going to be longitudinal in the micro-data sense. Before any analysis can be done, explicit account has to be taken of out-movers and in-movers so that they, and the truly continuous residents, can be treated appropriately.

The second problem is common to all micro-data sets, has to be solved by every analyst in a way that best fits his purposes, and is as follows. Even when it has been ascertained that the "same" family was indeed in the same dwelling unit both times, it may well be that the composition of that family has changed (slightly or drastically). A new child may be born or there may be a new family head, or a sub-family unit may have been created or destroyed. There are no obvious general rules about what changes require one to regard the changed unit as an essentially new one, but it is necessary to come up with some rule before the data can be properly used. The profession has not given much thought, hitherto, to the fact that a decision unit observed at time t may not exist at $t + 1$ or $t + 2$ (or may not have been there at $t - 1$). But when we attempt to use data generated by real units over a period of time such a problem is impossible to ignore. The solution, of course, depends on the conceptual foundations of one's specific analysis.

This, then, is the story of one relatively modest effort in the direction of a data bank. OEO aimed at producing a body of generally useful data (though focused on their concerns) and Census, Brookings, Assist, and the Poverty Institute have contributed in serial fashion to facilitating their use by researchers. It has taken a long while and we are still short of the goal.

Many of the problems appear to have been particular and specific to these data, but the order of magnitude of the problems, and the lack of any possibility of using previously-solved problems to expedite their solution, are common to all large micro-data bodies. And the person does not yet exist with the practical experience required to set up a data bank capable of handling such sets of data in their full generality. There are specialists who know about one specific applied concern, but their expertise is not yet transferable to other on-going data-collection efforts without a new learning process.

The Urban Graduated Work Incentive Experiment

The Graduated Work Incentive Experiment in New Jersey is a new departure in social experimentation which was funded in the summer of 1967 and fielded in August 1968. About 650 families (four sites in urban New Jersey and one in Pennsylvania) are receiving transfer payments of a negative income tax type, and roughly the same number are acting as a control group. The payments will continue over a three-year period. We collect income and family-size information for the experimental families every four weeks over the payment period, and during

this period both experimental and control families are administered an hour-long interview every three months.

Although payments started in 1968, it was not until the summer of 1970 that we were able to use our automated data system to retrieve any data, and the lag between when the information was coming in from the field and when it was retrievable by researchers was on the order of eight months. Since that time the lag has been becoming shorter and shorter, and data are now retrievable that are only three or four months out of the field.

As this short description will indicate, the "data-banking" problems faced in New Jersey are quite distinct from those faced in connection with the SEO. There are relatively fewer units of observation, but the information on each is voluminous. First of all there is information from thirteen hour-long interviews over the three-year payment period. These interviews have the same fifteen-minute core section (on labor supply) each time, but the rest of the hour is taken up with questions that vary from interview to interview. Some of the variables are measured repeatedly and some only once. Some of the families get lost—cannot be found or refuse to be interviewed—and most of the families undergo a change in composition or residence or both during the period of observation. The questionnaire structure (skip patterns and questions asked of different family members) is complex, imposing stringent standards on interview administration and completeness and consistency checking. In addition, there are four-weekly records of income and experimental payments for the part of the sample receiving "treatments."

Our aim is to produce a data source which is readily usable by research personnel. We are, therefore, concerned that an analyst be able to draw freely on variables from different survey waves or from other sources in order to "compose" and analyze a simple rectangular array of data for any sample of decision units that he may want to examine. This sounds like a modest goal, particularly since the same organization is responsible both for collecting and "banking" the data in this case. But no matter how simple and ultimately feasible this task may be, we can only proceed with frustrating slowness. There is no fund of experience to draw on in designing and executing the kind of data system we need—partly because the nature of the sample and study design are both novel and partly also because the technology, soft and hard, has been changing so rapidly.

Choice of Technology

As far as technology is concerned, I regard it as important to make an early and resolute decision about the kind of equipment and systems to be used. The choice should be made only among those alternatives that are already in sufficiently wide use to ensure (1) that valid information can be obtained on their performances in comparable applications *and* (2) that they have evolved a relatively stable, bug-free, and optimal set of software systems.

It requires some determination to avoid the choice of the latest equipment on the frontier. Such machinery always offers an exciting challenge to the system- and program-development staff, and the promise is always held out that the system eventually evolved will be superior to the potential of the more proven hardware. But I would emphasize that the costs of unforeseen difficulties and

delays are almost always very great. If the aim is to produce research in a reasonable period of time, the temptation to pioneer in computer systems must be resisted. Clearly one choice cannot be made for all time; but the strategy of *first* getting a working data facility and *then* catching up with the technology is the more prudent if delivering research products along the way is of any importance.

More needs to be said about how any micro-data system can, in the current stage of development, be ideally used by a researcher. The speed and cost of executing a given task of data manipulation is important in determining how much calculation will have to be done, and this may work in a somewhat perverse way: i.e., the slower and more costly it is to make one pass of the data file, the more likely it is that a researcher will try to anticipate all his potential needs on one pass. This strategy can be only partially successful in reducing future requests, but it does have a dramatic effect on the size of individual requests and on the amount of output accumulated: the more the analyst can replace an exhaustive set of possible choices with a sequence of choices conditioned upon previous outcomes, the more unnecessary calculation and superfluous output can be avoided. Hence the system should be designed to encourage a sharply-focused approach, and discourage the random shots.

The "Data Technician"

Ideally, a data system would be so automatic, self-describing, and well documented, that a research analyst could determine whether (and if so, how) the data could be used for his problem, and be able to carry out the job without assistance. It may well be possible to specify and design such a system, and it is certainly tempting to try and find one. But such an effort, again, will divert attention and resources away from getting research done in the near future. The more feasible approach for the next several years is to use a human intermediary who might be called a "data technician." The essential qualifications for such a person are: (1) the ability to communicate effectively with the researchers on the one hand, and with the computer technicians (operators, programmers, and system managers) on the other, and (2) a taste for detail that facilitates acquiring and retaining all of the "unwritten documentation," which seems to be an absolute requirement if one is to be able to use existing bodies of micro-data. To these might be added the third requirement—the capacity not to be easily discouraged.

There is now and for the foreseeable future a substantial fixed cost attached to the "first usage" of a new data set. Without the data specialist described above, who has become familiar with the data by struggling through that first use, much of that cost has to be incurred again by every subsequent user. Such a data technician can work directly with all users, determining first whether and generally how the data can be used to fill the researcher's need, and secondly whether to carry out the work him or herself or to train the user to do the job. This latter choice will depend on the size and complexity of the job and on the user's ability to learn enough to do it (or alternatively to pay for the service of having it done). But without such a person (who is either familiar with the data or has the responsibility for becoming so), users will be scared off a new data file by the complexity of "getting into" it. And those who are not put off immediately will become dis-

couraged (or impoverished) to the point of abandoning the effort before they get any results. To repeat: Such a data-specialist could be dispensed with in an ideal "data bank," but for the foreseeable future I believe it to be an indispensable part of any organization that aims at facilitating the use of complex micro-data sources.

The Use of Still-Accumulating Data Files

Additional problems and opportunities are encountered when a body of data is being used while still in the process of collection—as is the case with the Negative Income Tax data being gathered in New Jersey. Such was (and is) the need for any information on this subject that the data system had to become operational before the eventual dimensions of the data base were fixed. Research production and the programming related to it, therefore, compete for time and budget with the development of the data system *per se*. Files extracted for analytic use will become obsolete as errors are corrected, coding is improved, data are added, and temporarily lost families recovered. Early results must, therefore, be expected to be inconsistent (usually in trivial ways) with those obtained later in the process.

Important offsetting advantages do, however, exist. The fact that data producers, data users, and system designers *have* to work together reduces the chance of serious mistakes—those requiring part of the basic job to be done over again. Interim or preliminary use of the data results in the discovery of problems and ambiguities in time for revisions, before the difficulty has been replicated throughout the data. In retrospect, for example, it is quite clear that the SEO would have been available in useful form much earlier (and would in fact have been a superior data set) if there had been serious and urgent analytic interest at the Census Bureau within the group responsible for producing the research-ready tape.

Summary and Advice

Despite the many recent technical developments in computer hardware and software systems I remain awed at the difficulty of building a usable data bank, and also awed at the readiness with which such a concept is often discussed as feasible. My own experience suggests that efforts in this direction err on the over-ambitious side, and consequently are so long in gestation that the interest of the research analyst is lost. I cannot overemphasize this: If the primary objective is to facilitate real research, start small and develop competence with one basic body of data. Once you have handled that task, proceed to others.

The very latest in technology is another pitfall to be avoided. Unless you have endless time, money, and patience, use equipment and software systems that have known and stable characteristics. The newest and fastest may eventually be the best, but getting it to work will *always* take longer than anyone expects.

At the present time, a person-plus-machine system, utilizing what I have called a data technician as a communicator and ambulating documentation file, is the best way to get started. Again, it may be that a more direct system can evolve from this, as the technician finds ways to reduce the number of simple and repetitive requests. But there is simply not enough experience in this area at this time

for anyone to feel confident about starting out with an automated system alone.

Finally, I would urge that a data bank be focused from the start on the needs of specific analysts—people who exist, are alive, and on the premises. They must be persuaded to become involved in the process of system design from the start; and they must be impatient enough for results to try out and test pieces of the system and the data file as soon as they begin to take shape.

All this may sound like a counsel of despair. That is not my intent. But, however ambitious one wants to be in planning toward some ultimate general data bank, it is imperative to start somewhere and get some real work *done*. The beginning must be quite modest if we are to make any progress at all.